SYNOPSYS®
Silicon to Software™

Newsletter for semiconductor process and device engineers

December 2016

TCAD news

Latest Edition

Welcome to the IEDM 2016 edition of the TCAD News. The IEDM is always a special event for our TCAD team as it represents the continuous innovation and dynamism of the semiconductor industry which we strive to support. With 7nm node early production set to start in 2017, and 5nm node R&D well underway, leading edge logic manufacturers continue to deliver more performance and higher density at reduced power. Yet, the cost and technical challenges of developing these advanced technologies is rising exponentially. To address these challenges, design-technology co-optimization (DTCO) has become essential for selecting process technology options using design-level criteria. In 2016, Synopsys developed and introduced to the market DTCO simulation flows to enable our customers to make better and faster decisions on which process technology options to adopt in order to meet performance, power and area (PPA) targets.

This newsletter describes an important aspect of DTCO: the design and optimization of new transistor and interconnect architectures using circuit-level metrics derived from the simulation of test circuits, enabled through the extraction of SPICE models from TCAD data made possible through our acquisition of Gold Standard Simulations (GSS) in May, 2016.

The second article highlights a new capability in Sentaurus Device to simulate negative-bias temperature instability (NBTI), a key reliability concern in advanced CMOS technologies. This work is the result of a collaboration between Synopsys and IIT-Bombay.

With the approaching holiday season, I would also like to take this opportunity to wish you happy holidays and a prosperous New Year.

With warm regards,

Terry Ma
VP Engineering, TCAD

Contact TCAD
For further information and inquiries:
tcad_team@synopsys.com

Synopsys Solutions for the Development of Advanced Logic Process Nodes

Industry Challenges at Advanced Nodes Motivate Design-Technology Co-Optimization (DTCO)

For over 35 years, the semiconductor industry has delivered transistor scaling defined by Moore's law through the optimization of the silicon planar MOSFET. Although many significant innovations had to occur along the way, such as copper interconnects, strained silicon, immersion lithography, and multi-patterning, to name a few, scaling progressed along a path that allowed semiconductor manufacturers to predictively deliver new process nodes every 18-24 months.

The introduction of FinFET devices at the 14nm process node, however, brought about a significant increase in the complexity of technology development due to the FinFET's inherent 3D device structure. The 3D structures led to new process integration challenges that did not exist with traditional planar MOSFET structures. These challenges have only become more acute with the subsequent scaling of the FinFET at the 10nm and 7nm process nodes. With the current research and development of the 5nm node and beyond, new challenges emerge, particularly in the assessment and integration of novel materials and transistor architectures to improve performance, and the evaluation of novel lithography techniques to maintain scaling. Along with this rise in process complexity, the smaller geometries have made transistor and interconnect performance more interdependent and susceptible to variability.

From the point of view of semiconductor manufacturers, these challenges manifest themselves in the need to evaluate more process options, materials, and transistor architectures, concurrently with the primary design rules to enable their implementation. The down-selection of these options must be based on not just single transistors, but also design-level criteria such as the power, performance, area (PPA) of logic constructs representative of the IC products targeted by the process node. The major implication to semiconductor manufacturers of this new paradigm is that unless new methodologies are implemented, technology development timelines will suffer and the risk of making the wrong technology choices will rise. Methodologies to support his new development paradigm are known as Design-Technology Co-Optimization (DTCO).

In This Edition

An evident way to maintain the development timelines is to begin the evaluation of the technology options earlier in the development cycle. This “shift left” in time needs to be almost entirely based on simulation since wafer data is often non-existent, particularly in the case of new materials. Therefore, from the simulation point of view, “shift left” calls for DTCO enablement in the pre-wafer phases of technology development, at one end reaching deeper into materials modeling at atomistic scales, to enable the evaluation of novel materials, and at the other end linking with the simulation of design-level circuits.

To address these challenges, Synopsys has developed a pre-wafer DTCO flow with the objective of enabling simulation-based narrowing down of process and device options, reducing expensive and time consuming wafer-based process and design iterations, and ultimately leading to the delivery of more competitive products. A key objective of this flow is the creation of a high quality early Process Design Kit (PDK) to make possible the design-level activities of DTCO, such as the early design and evaluation of IP and test circuits.

DTCO is a multi-disciplinary activity involving process integration, lithography development, TCAD simulation, IP design and design functions. In this article, we describe three key TCAD contributions to DTCO flows: design advanced transistor architectures, process emulation and interconnect simulation, and extraction of SPICE models from TCAD data. A companion article in this newsletter discusses the important topic of negative-bias temperature instability (NBTI), a significant reliability challenge in advanced transistor architectures.

Design of Advanced Transistor Architectures

Research and development of new transistor architectures to replace the FinFET is underway. Options include vertical and horizontal nanowire FETs, nanosheet FETs, and tunnel FETs. A common theme among these future transistor architectures is the critical importance of designing the transistor and lower interconnect layers concurrently, as interconnect parasitics increasingly limit performance gains. Examples are the rising contact to gate capacitance and resistivity of the contacts and middle-of-line (MOL) due to surface scattering in narrow wires and vias. Performance and power evaluations of candidate transistor architectures can be done with ring oscillators made up of inverter or NAND gates, with the cells incorporating the lower interconnect layers.

A key objective for new process nodes is power reduction; from the point of view of transistor design this translates to more effective suppression of short-channel effects. At the 5nm node, gate-all-around nanowires and nanosheets are being investigated as promising candidates owing to their superior electrostatics relative to FinFETs. The small cross-sectional size of the nanowires and nanosheets introduces significant 2D quantization effects which require the solution of the Schrödinger equation. On the other hand, transport along the direction of the current is quasi-ballistic due to the very short gate lengths. These phenomena can be modeled with the subband Boltzmann transport equation coupled with the solution of the 2D quantization problem on the cross-section. This capability is available in Sentaurus Device QTX and is illustrated here for nanowire structures consistent with 5nm node rules. An accompanying application note with more details is available through Solvnet [1].

Silicon is used as the channel material. Modifications to the band structure are effected through an ellipsoidal model calibrated to a two-band model for electrons and a six-band model for holes. Phonon scattering for electrons comprises three g-type and three f-type intervalley inelastic processes and one intravalley acoustic scattering, whereas for holes, optical phonon scattering and acoustic phonon scattering are included. Surface roughness scattering for electrons and holes use an exponential model of the power-spectrum density function. Lumped series resistance models the source/drain contact resistance and the source/drain (S/D) epi diffused resistance.

Several slices are defined across the channel to allow the solution of the 2D Schrödinger equation on each slice. The 1D subband-BTE is solved in the channel direction, and the Poisson equation in the 3D nanowire structure. All equations are solved self-consistently.

The transport direction is defined along the z-axis (set to the <110> crystal orientation in the project) and the confinement directions are aligned to the x- and y-axes. The nanowire structure is shown in Figure 1.

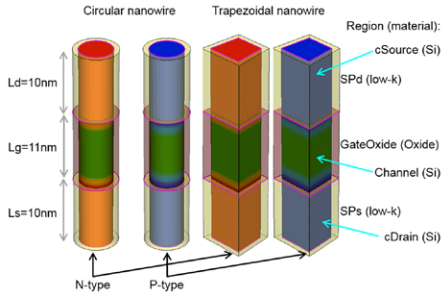


Figure 1. Silicon nanowire simulation structure consisting of circular and trapezoidal cross sections; gate oxide thickness is 0.8 nm, diameter of the circular nanowires is 5 nm, and geometric parameters of the trapezoidal nanowires are top width=4.5 nm, bottom width=5.5 nm, and height=5 nm.

The structure consists of a silicon nanowire, oxide surrounding the nanowire channel region, low-k in the extension region, and contacts. The channel length is 11 nm and the S/D extension length is 10 nm. The diameter of the circular nanowire is 5 nm. The top width, the bottom width, and the height of the trapezoidal nanowire are 4.5 nm, 5.5 nm, and 5 nm, respectively. The oxide thickness is 0.8 nm. The channel is undoped and the S/D extension doping is 10^{20} cm^{-3}. The source and drain doping slopes are generated by diffusion at 1100°C as shown in Figure 2.

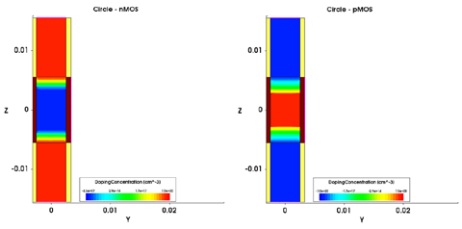


Figure 2. Doping profile in silicon nanowire simulation structure: (left) circular NMOS and (right) circular PMOS. Uniform doping of 10^{20} cm^{-3} is specified in the source and drain regions and 1100°C annealing is performed.

Figure 3 shows the mesh generated with Sentaurus Process. All the mesh points must be located on the slices where the 2D Schrödinger equation is solved. The simulation structure is created through the extrusion of a 2D cross-section mesh.

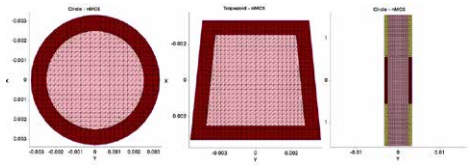


Figure 3. Meshes of silicon nanowire simulation structures: 0.2 nm mesh size at the interface between the gate oxide and the silicon channel, and 0.5 nm mesh size at the S/D junction along the channel direction.

When the cross section of the nanowire becomes smaller, the gap of the subband energy increases and fewer subbands will contribute to transport. Since the number of subbands determines the simulation time (the number of Boltzmann equations and the inter-subband scattering matrix), it is important to select the appropriate number of subbands to keep the simulation time reasonable.

The initial estimate of the number of subbands can be given by Sentaurus Band Structure analysis of the 2D cross section of a nanowire. With this initial estimate, a few simulation splits on the number of subbands will determine the optimum setting. However, if the application is the study of the cross section of a transistor, the subband number should vary accordingly to the size of the cross section. In the examples shown here, the number of subbands vary between 6 and 12 for n-type, and 18-24 for p-type.

Figure 4 shows the subband-BTE simulation results of Id–Vg curves for the circular and trapezoidal nanowires.

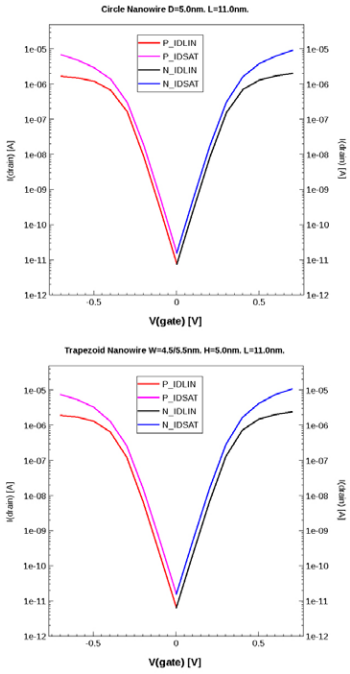


Figure 4. Id–Vg characteristics for (top) circular nanowires and (bottom) trapezoidal nanowires (n-type and p-type).

The on-current of the n-type nanowires is higher than that of the p-type nanowires. The trapezoidal nanowires provide a higher current than the circular nanowires due to their larger cross-section area. However, the trade-off to the subthreshold is observed to be minimal.

Figure 5 shows the current, electron density and density-of-states (DOS) of the n-type circular nanowire in the space of the channel coordinate and the energy.

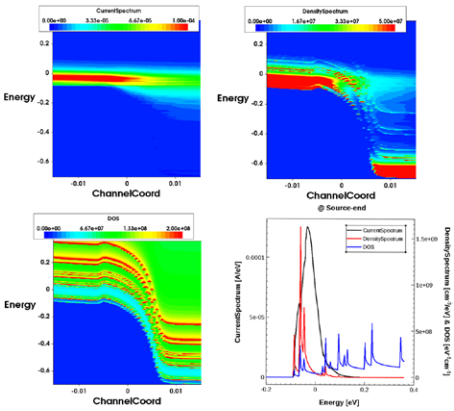


Figure 5. (upper left) Current, (upper right) density, and (lower left) density-of-states of the n-type circular nanowires along the channel direction in the energy space: Vg= 0.7 V, Vd= 0.7 V, and (lower right) cut at source end.

The source is shown to the left in the plot and the drain is shown to the right in the plot. The current spectrum shows that electrons are scattered in the drain and lose energy. The density spectrum shows the high electron concentration in the source and the drain. Figure 5 (lower right) shows the cut at the source end. It illustrates that the DOS follows the trend of $Energy^{-1/2}$, which can be derived analytically in the quantum wire 1D structure. The broader distribution of the current spectrum demonstrates the contribution of electrons with higher momentum.

The corresponding current, hole density and DOS of the p-type circular nanowire in the space of the channel coordinate and the energy is shown on Figure 6.

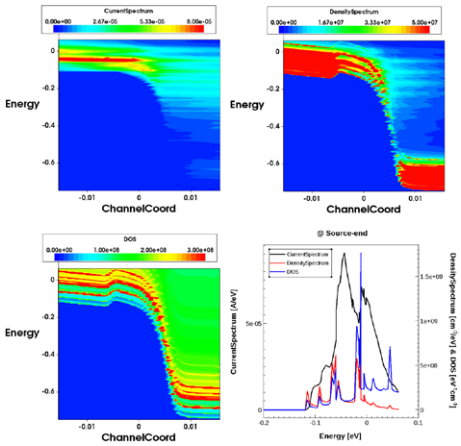


Figure 6. (upper left) Current, (upper right) density, and (lower left) density-of-states of the p-type circular nanowires along the channel direction in the energy space; Vg= –0.7 V, Vd= –0.7 V, and (lower right) cut at source end.

Figure 7 shows the distribution function of the circular nanowires in log scale. For the n-type nanowire, the distribution function is shown for the 0th subband of the Δ1-valley in the linear and saturation modes. For the p-type nanowire, the distribution function is displayed for the 0th subband of the Γ- valley.

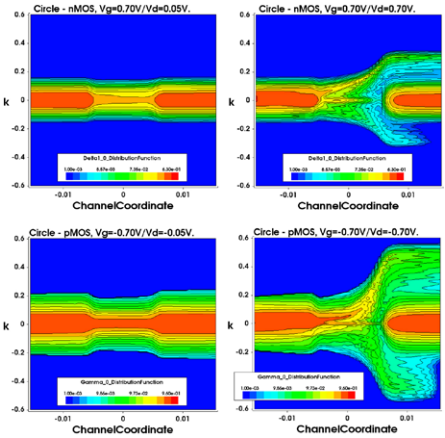


Figure 7. Distribution functions (in log scale) of the circular nanowires (0th subband of Δ1-valley for n-type nanowire, and 0th subband of Γ-valley for p-type nanowire: |Vg|= 0.7 V, |Vd|= 0.05 V/0.7 V.

The distribution function of the linear mode is close to the equilibrium Fermi function. However, the saturation mode distribution function shows the quasi-ballistic transport as well. Noticeable ballistic transport is observed in the channel (see Figure 7 (upper right and lower right)). Some carriers lose energy due to the optical phonon scattering near the drain, and a few carriers return to the source.

Such detailed modeling of transistor architectures available in Sentaurus is instrumental for the exploration of viable candidates and the eventual optimization of the candidate structures. However, the transistor architecture must be evaluated together with the associated interconnect parasitics since performance trade-offs often emerge between transistor and interconnect design.

Process Emulation and Interconnect Parasitic Extraction

Concurrent evaluation of interconnect performance based on the same rules used to define the transistor architecture is achieved through process emulation. Process Explorer inputs the mask layout and process flow to build 3D representations of the technology. Its fast 3D modeling, combining geometric topographical steps with physical etch and deposition models, allows iterative exploration of design rules and process module options. The resulting 3D structure is then simulated with Raphael, the gold-standard interconnect solver, to extract the resistance and capacitance of the structure.

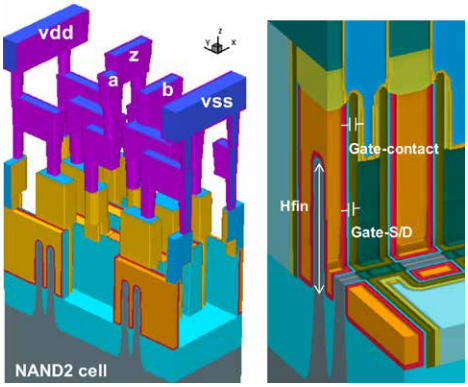


Figure 8. FEOL and MOL 3D structure for a FinFET-based NAND cell generated with Process Explorer.

Table I shows the parasitic capacitances for this structure simulated with Raphael for two fin heights of 45nm and 30nm. The higher parasitic capacitance incurred by the taller fin must be traded-off against its capability to deliver higher drive current. Such trade-offs can be evaluated at the circuit level if a SPICE model is available for the transistor. This is topic of the next section.

Capacitance	net1	net2	Hfin45nm	Hfin30nm	delta(45vs30)
C_0_1	vss	vdd	8.46E-19	7.39E-19	14%
C_0_2	vss	b	1.23E-17	1.16E-17	6%
C_0_3	vss	a	4.24E-17	3.42E-17	24%
C_0_4	vss	ni	1.05E-17	1.01E-17	4%
C_0_5	vss	z	4.42E-17	3.65E-17	21%
C_1_2	vdd	b	4.40E-17	3.54E-17	25%
C_1_3	vdd	a	4.51E-17	3.67E-17	23%
C_1_4	vdd	ni	3.52E-19	3.20E-19	10%
C_1_5	vdd	z	2.88E-17	2.81E-17	2%
C_2_3	b	a	1.92E-17	1.70E-17	13%
C_2_4	b	ni	4.28E-17	3.52E-17	22%
C_2_5	b	z	9.11E-17	7.56E-17	21%
C_3_4	a	ni	4.03E-17	3.27E-17	23%
C_3_5	a	z	4.64E-17	3.88E-17	20%
Increase in fin height from 30nm to 45nm results in coupling capacitance changes, ranging from 3% to 25%.					

Table I. Parasitic capacitances of FinFET-based NAND cell for fin heights of 45nm and 30nm simulated with Raphael.

TCAD-SPICE Link to Enable Early Simulation of Test Circuits

This section describes a methodology for extracting SPICE models from TCAD data. This methodology enables the early simulation of test circuits, such as ring oscillators (RO), library cells and SRAM cells, as part of the design-level assessment for DTCO.

The SPICE model extraction is based on a hierarchical approach which includes nominal model, response surface model and statistical model extraction. One of the distinctive outputs of this methodology is the generation of models that capture non-Gaussian distributions of key transistor figures of merit and their correlations. Correlations between process and statistical variability are preserved.

The target data for the SPICE model extraction is generated by TCAD. For nominal model extraction, Sentaurus Device simulates current-voltage (I-V) and capacitance-voltage (C-V) curves consistent with the requirements of the specific SPICE model and its underlying extraction strategy. The Sentaurus Device simulations, and its predecessor Sentaurus Process simulations, are carried out for device structure splits to capture the variability arising from geometric variations such as fin height, fin width, nanowire diameter, gate length, etc. With each split assigned as a factor in a design of experiments (DoE), the Sentaurus Device output is then fitted with a response surface model. For the geometric splits defined, the resulting RSM is typically smooth and therefore can accurately model intermediate points within the process window defined by the DoE points.

Local variability, due to gate edge roughness, random doping and trap fluctuations, and metal gate granularity can be added at each of the DoE points. However, the output is now generated by the variability engine GARAND in view of its unique robustness and speed for these types of simulations. Upon calibrating GARAND to Sentaurus Device across the DoE points, an operation that is automated, GARAND then generates hundreds (typically 100-500) of microscopically different devices subject to statistical variation. The data output from these statistical simulations becomes the target data for statistical SPICE model extraction. Figure 9 shows the statistical target Id-Vg data at three DoE points for a 14nm SOI FinFET technology using the BSIM-CMG SPICE model.

In the afore-mentioned flow, the DoE and the generation of the full set of target I–V characteristics (nominal and statistical) are fully automated using the automation tool Enigma. This automation provided by Enigma is essential for handling the large amount of data, input files and simulation jobs involved in the methodology. The target data are harvested in a common database and automatically accessed by the SPICE model extractor Mystic.

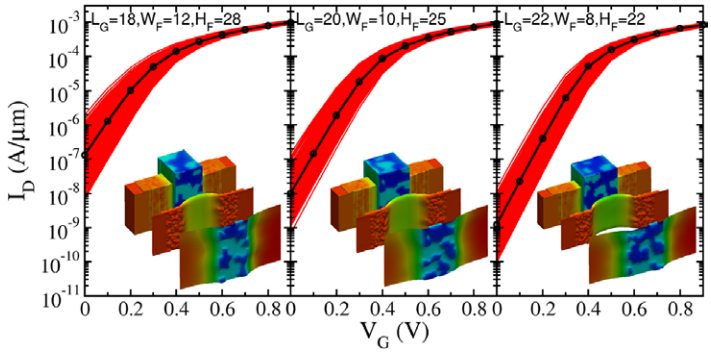


Figure 9. Statistical target ID–VG characteristics at three nodes of the DoE space. Inset: microscopically different devices from the sample, within which the front slice shows the potential across the gate oxide modulated by MGG and the middle slice shows the electron density along the mid-channel.

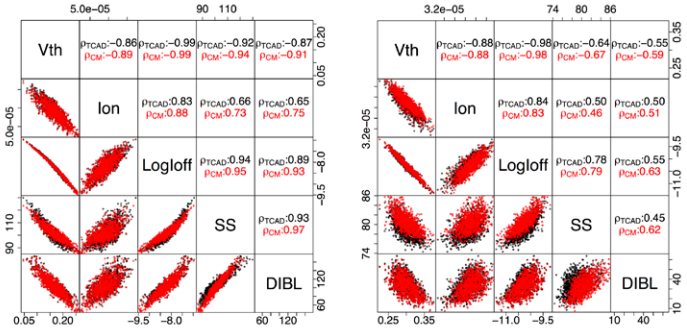


Figure 10. Distributions and correlations between the essential figures of merit obtained from the TCAD simulation and the response surface compact macromodel at two extreme DoE corners. (a) LG = 18 nm, WFIN = 12 nm, and HFIN = 28 nm. (b) LG = 22 nm, WFIN = 8 nm, and HFIN = 22 nm.

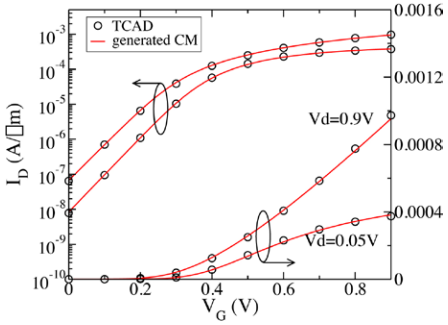


Figure 11. Comparison between between SPICE model generated and TCAD-simulated generated and TCAD-simulated transistor characteristics with a random set of LG (=18.5 nm) and WFIN (=11.5 nm) with HFIN = 25 nm.

The extraction strategy is adjusted for different technologies and SPICE models. Careful selection of the SPICE extraction parameters is required in order to capture the variability effects and preserve their correlation.

Figure 10 shows the distributions and the correlations among the essential figures of merit obtained from the TCAD atomistic simulations and from the RSM model for two DoE points.

Statistical SPICE models are generated with RandomSpice. The ModelGen technology in RandomSpice is designed to account for non-Gaussian parameter distributions. The resulting statistical SPICE models generated with RandomSpice are used to overcome the limitations with using finite input samples.

Subsampling errors, which artificially distorting the output distribution and are a concern when sampling rare events, are avoided. Another feature of ModelGen is to allow the variation captured by the RSM and statistical simulations to be continuously modeled across the entire DoE space.

Simpler methods, such as principal component analysis (PCA) and nonlinear power method (NPM), are limited by assumptions of Gaussian marginal distributions, in the case of PCA, and by numerical stability, in the case of NPM.

Figure 11 shows the excellent fit between TCAD and SPICE model generated I-V curves for two randomly selected points within the DoE space.

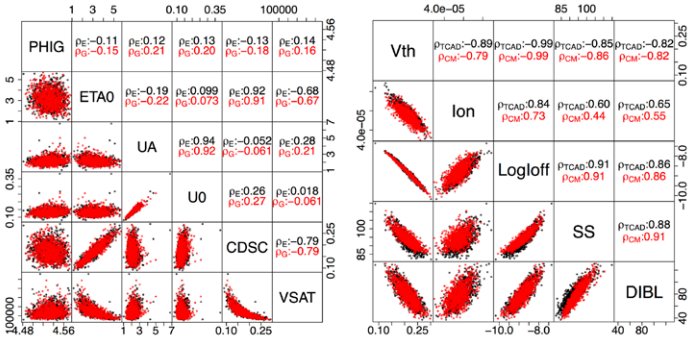


Figure 12. Left) Extracted (E) and generated (G) distributions of the LV CM parameters at the randomly selected LG and WFIN. (right) Distribution and the correlations of the key FinFET figures of merit extracted from the statistically generated CM at the randomly selected LG and WFIN and the corresponding TCAD simulations.

Figure 12 shows the distribution and the correlations of the key FinFET figures of merit extracted from the statistically generated SPICE models at the randomly selected DoE points and the corresponding TCAD simulations. The excellent correlations shown are an indication of the precision with which the statistical SPICE model extraction reproduces the statistical target data and its correlations.

The TCAD-based variability-aware SPICE model extraction described here has the requisite automation to allow systematic DTCO investigations. One key aspect of this methodology is that it generates statistical SPICE models that capture the interplay between local and global variability. When combined with TCAD simulation of candidate transistor architectures and interconnect stacks, this methodology is very useful in the circuit evaluation of SRAM, ROs and other circuits that are representative of the PPA metrics targeted by the new process node.

References

[1] "Three-Dimensional Simulation of Silicon 5nm Node Nanowires using the Subband Boltzmann Transport Equation," Synopsys Solvnet. https://solvnet.synopsys.com/retrieve/2521688.html

[2] X. Wang, B. Cheng, D. Reid, A. Pender, P. Asenov, C. Millar, and A. Asenov, "FinFET Centric Variability-Aware Compact Model Extraction and Generation Technology Supporting DTCO," IEEE Trans. on Electron Devices, Vol. 62, No. 10, Oct 2015, pp. 3139-3146.

NBTI Modeling in Sentaurus Device

Subrat Mishra, Narendra Parihar, Rakesh Rao, and Souvik Mahapatra, IIT Bombay, Mumbai, India
Hiu Yung Wong, Steve Motzny, and Victor Moroz, Synopsys Inc., Mountain View, CA, USA

Introduction

Negative Bias Temperature Instability (NBTI) is a crucial and well-known p-MOSFETs reliability concern. It became an issue for planar devices with the advent of Silicon Oxynitride (SiON) gate insulator technology, and continues to impact the stability of High-K Metal Gate (HKMG) FinFETs. NBTI is due to gradual buildup of positive charges in the gate insulator, and results in the degradation of device parameters in time, such as threshold voltage (VT), transconductance (gm), sub-threshold slope (S), linear (IDLIN) and saturation (IDSAT) drain current, etc [1].

Figure 1 shows the schematic of a HKMG stack, consisting of SiO_2 interlayer and HfO_2 High-K layer sandwiched between the Si channel and TiN gate. After several years of debate, it is now well accepted that the positive gate insulator charges primarily come from two mutually uncorrelated sub-components: interface trap generation and hole trapping [1]. Interface traps are created by breaking of H passivated defects at the Si/SiO_2 interface and inside the gate insulator bulk (effectively considered at the SiO_2/HfO_2 interface). Hole trapping takes place in pre-existing process-related defects in the SiO_2 interlayer. For harsher stress conditions (e.g., high voltage and/or high temperature stress), generation of bulk insulator traps (E' centers) also contributes to overall positive gate insulator charges [1].

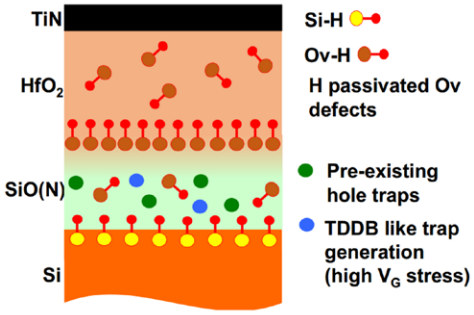


Figure 1. Schematic showing the HKMG gate insulator stack and precursors for trap generation and trapping that governs overall NBTI.

In this article, implementation of the two primary NBTI sub-components in Sentaurus Device is discussed, and simulation results are shown for planar, FinFETs and GAA NWFETs. It is now possible to use TCAD for co-optimization of power, performance and NBTI reliability for advanced technology nodes.

TCAD Simulation Setup

Figure 2 shows the simulation setup. Interface trap generation and passivation (ΔNIT) respectively during and after NBTI stress is simulated using the Reaction-Diffusion (RD) model [1], implemented in Sentaurus Device using the Multi State Configuration (MSC) – Hydrogen transport degradation framework [2]. RD model simulates density of traps; however, it is the charge occupancy of these traps that impacts device degradation. Therefore, ΔNIT simulated with the RD model is augmented by the Transient Trap Occupancy Model (TTOM) [3], to calculate the contribution of charged traps (ΔVIT) to threshold voltage shift (ΔVT). Hole trapping and de-trapping (ΔVHT) are calculated using the extended Nonradiative Multi-Phonon (eNMP) model [4], also incorporated in Sentaurus Device. Overall ΔVT is the summation of ΔVIT and ΔVHT.

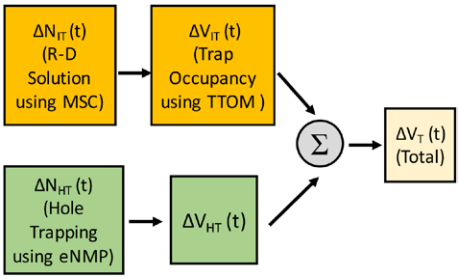


Figure 2. Sentaurus TCAD framework for modelling NBTI ΔVT. The interface trap generation (ΔVIT) component is obtained from RD solution using MSC degradation model and augmented by TTOM for trap occupancy. The hole trapping (ΔVHT) component is obtained from extended Nonradiative Multi-Phonon (eNMP) model.

Simulation of Trap Generation

The double-interface H/H_2 RD model [1] is implemented to simulate generation and passivation of traps during and after NBTI stress. Figure 3 shows the model as applicable for a HKMG stack. During stress, inversion layer holes get captured by H passivated defects at the Si/SiO_2 interface, and the weakened bonds are subsequently broken by thermal excitation. The released H atoms diffuse and subsequently react with H passivated defects in the gate insulator bulk (lumped at the SiO_2/HfO_2 interface for ease of implementation) and release H_2 molecules, which diffuse further into the gate stack.

During recovery, H_2 molecules diffuse back and passivate broken bonds at the SiO_2/HfO_2 interface and release H atoms, which further diffuse and passivate broken bonds at the Si/SiO_2 interface.

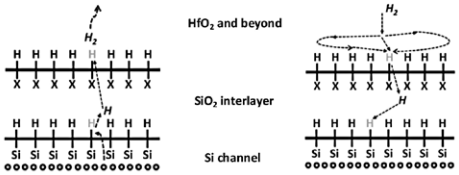


Figure 3. Schematic of double interface H/H_2 RD model for trap generation during stress (left) and trap passivation during recovery after stress (right).

The Multi State Configuration (MSC) - Hydrogen Transport degradation model in Sentaurus Device [2] is used for self consistent solution of the Poisson equation and H_2 transport equations, with quantum corrected inversion layer hole density, to simulate trap generation and passivation during and after NBTI stress. Figure 4 (left) depicts the state diagram of the MSC model, and shows the initial and final states as well as the forward and reverse reaction rates used to implement the double interface H/H_2 RD model. The equations describe breaking and passivation of bonds at the Si/SiO_2 and $SiO2/HfO_2$ interfaces of the HKMG stack, and diffusion of H and H_2 species in the HKMG stack and beyond.

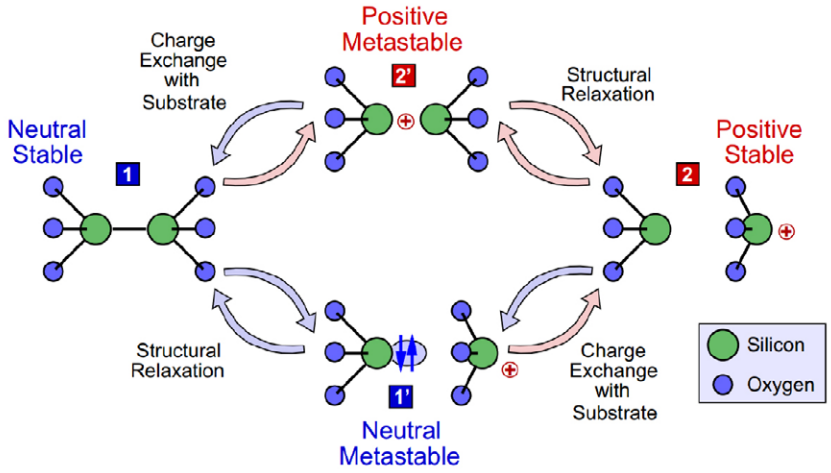


Figure 5. State diagram for eNMP model showing different states and transition between states. S1 = Neutral stable state; S2 = Positive stable state; S1' = Neutral metastable state and S2' = Positive metastable state.

During stress, H atoms have to "find" H passivated bonds for reaction at the $SiO2/HfO_2$ interface. During recovery, H_2 molecules and H atoms have to find broken bonds for reverse-reaction respectively at the $SiO2/HfO_2$ and Si/SiO2 interfaces. Since H atoms diffuse quite fast, they can find a bond to react very quickly during stress or recovery. However, due to slow diffusivity, H_2 molecules have to hop near the $SiO2/HfO_2$ interface before broken bonds are found for reverse reaction, see Figure 3 (right). This would slow down the recovery with the passage of recovery time, as unpassivated bonds reduce in density and are difficult to find, refer to [1] for detailed discussion. Such a stochastic "hopping" effect is incorporated in the continuum simulation framework by slowing down H_2 molecule diffusivity in time only during recovery, by using a physical model interface (PMI) implementation.

Moreover, while the RD model simulates generation and passivation of traps, it is the charge occupancy of these traps that determines gate insulator charges during and after NBTI stress. Figure 4 (right) shows the energy band diagrams of a HKMG stack for stress and recovery. During recovery, some of the generated traps would go below the Fermi level, capture electrons and become neutral. Therefore, these traps would not contribute to positive gate insulator charges, even if the traps exist physically. RD model is augmented by TTOM [3] to calculate density of active generated interface traps during and after NBTI stress.

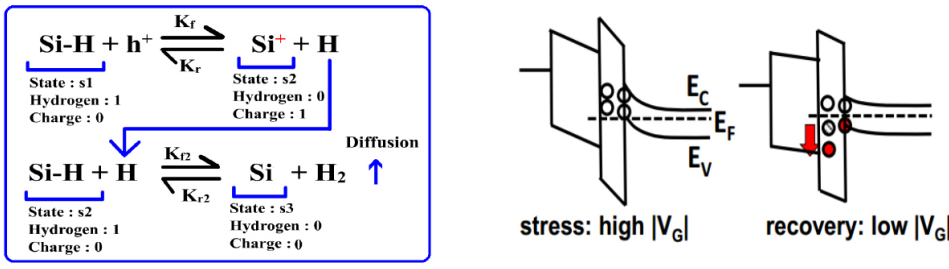


Figure 4. State diagram for reactions involved during NBTI based on double interface H/H2 RD model implemented in Sentaurus Device using MSC degradation framework (left). Energy band diagrams during stress and recovery showing the need for capturing trap occupancy (using TTOM) during recovery (right).

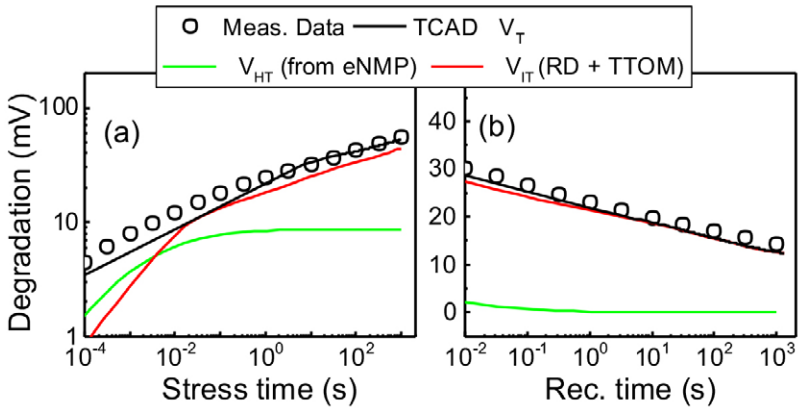


Figure 6. Measured (symbols) and TCAD simulated (lines) degradation using the framework of Figure 2 during and after NBTI stress. The individually simulated sub-components are also shown.

Simulation of Hole Trapping

The extended Nonradiative Multi-Phonon (eNMP) model [4] is incorporated in Sentaurus Device for calculating hole trapping in pre-existing traps [2]. Figure 5 describes the different states associated with the eNMP model. During stress, transition occurs from the neutral state (S1) to the final trapped state (S2) via an intermediate metastable state (S2'), and the time-dependent occupancy of S2' and S2 provides the kinetics of hole trapping. During recovery, transition occurs from the final S2 state to the neutral S1 state, either via the metastable state S2' if recovery bias is high, or via the metastable state S1' for low recovery bias. The reactions between S1 and S2' and S2 and S1' are handled by non-radiative multiphonon process. The reactions between S2' and S2 and S1' and S1 are handled by thermionic process. Refer to [4] for further details.

Prediction of Measured Data

Figure 6 shows the TCAD model prediction (lines) of measured NBTI degradation (symbols) in a planar MOSFET during and after NBTI stress. The underlying trap generation and hole trapping related sub-components are also shown. During stress, ΔVHT saturates fast and has much lower relative contribution on long time ΔVT. On the other hand, ΔVIT keeps on evolving in time and dominates long time ΔVT. Therefore, ΔVIT impacts NBTI degradation at end of life, which is especially true for well-optimized production quality devices [5]. During recovery, ΔVHT recovers relatively fast, while the reduction in ΔVIT due to trap neutralization and trap passivation continues for long time.

Figure 7 (top) shows the prediction (lines) of measured stress data (symbols) at different stress bias and temperature for the device of Figure 6, while Figure 7 (bottom) shows the corresponding prediction (lines) of measured recovery data (symbols) for this device, for different stress bias and stress time. Identical model parameters are used to simulate experimental results in Figures. 6 and 7. The framework can predict NBTI stress and recovery for wide variety of experimental conditions. In particular, long-time experimental data can be successfully predicted, which is useful for the determination of end-of-life.

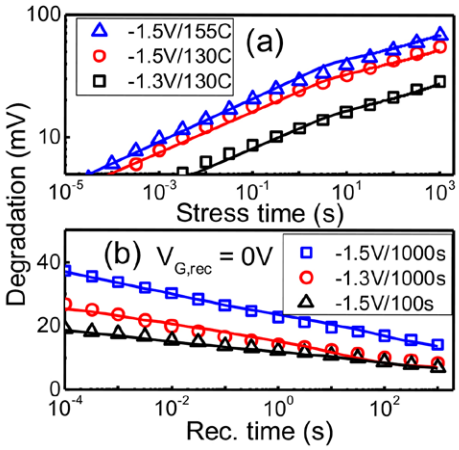


Figure 7. Measured (symbols) and TCAD simulated (lines) degradation using the framework of Figure 2 during (top) stress at different bias and temperature and (bottom) recovery at different stress bias and stress time.

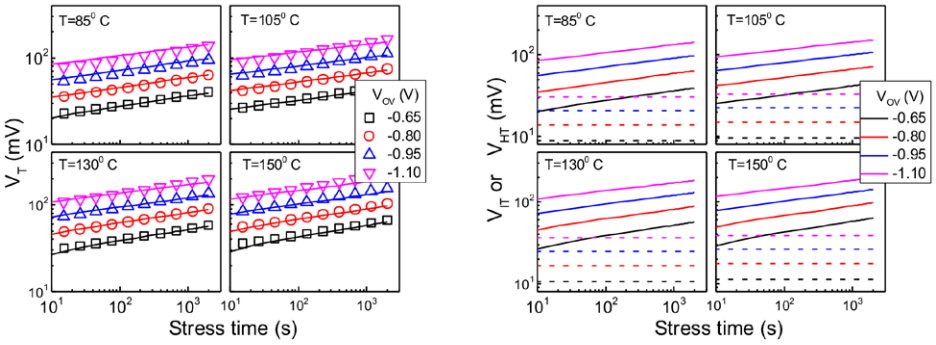


Figure 8. (Left) Simulation of NBTI degradation (lines) and measured bulk FinFET degradation data (symbols) for four different stress biases and temperatures. All different experimental conditions are predicted using identical model parameters. (right) Simulated interface trap generation (solid lines) and hole trapping (dashed lines) subcomponents corresponding to the left figure panels.

Figure 8 (left) shows shows the prediction (lines) of measured stress data (symbols) at different stress bias and temperature for a FinFET device. The framework can predict such wide variations of experimental conditions using a consistent set of parameters. Figure 8 (right) shows the corrsponding simulated undelying trap generation (solid lines) and hole trapping (dashed lines) sub-components. Hole trapping saturates at long time. Trap generation continues to build up in time and dominates overall NBTI at long time. Therefore, trap generation primarily determines NBTI degradation at end of life.

Architecture Dependence

Figure 9 (left) shows the isometric view of the devices used to demonstrate the capability of the TCAD NBTI framework [6]. Only trap generation and trap occupancy are simulated, as hole trapping has been found to be negligible in production quality gate stacks [5]. 3-D simulations were carried out using realistic Back End of Line (BEOL) structure for the FinFETs and NWFETs to facilitate diffusion of H_2 (Figure 8 (right)). The Low-*k* material surrounding the gate metal vias for isolation is not shown in the figure for better viewing. In these dimensions, inversion carriers experience strong electrostatic and geometrical quantum confinement effects. Therefore, Poisson equation is solved fully coupled with the density gradient quantum corrections, as well as hole and hydrogen transport equations. Dirichlet boundary conditions for [H] and [H_2] are assumed at the contacts, i.e., H and H_2 diffuse through the STI and vias and finally get absorbed at the contacts. The BEOL simulation domain is kept sufficiently long in order not to encounter boundary absorption related artifacts, see [6] for details. All model parameters are calibrated against FinFET degradation data for different stress bias and temperature (Figure 8).

It is important to remark that all the reaction and diffusion related RD model parameters of the MSC Hydrogen transport framework, except those associated with the forward reaction rate responsible for breaking of Si-H bonds at the Si/SiO2 interface, remain constant for different gate stacks. Only 3 gate stack dependent parameters are needed for the forward reaction term (pre-factor, field acceleration factor and temperature activation energy), which makes model calibration quite straightforward. Refer to [1], [3], [6] for additional details.

NBTI stress and recovery is simulated on planar MOSFET, bulk FinFET and GAA NWFET. Figure 10 shows (a) time evolution of ΔVT (=ΔVIT) for stress and recovery, normalized to end-of-stress data (b) longer-time power-law time exponent n for stress, obtained using linear regression between 100s-1000s and (c) fractional recovery, simulated for different device architectures. Simulated ΔVT shows similar stress and recovery temporal dependence for different device architectures. However, the magnitude of ΔVT is different for different device architectures at constant overdrive (VOV) stress, due to different electrostatics and confinement effects that causes a difference in the forward reaction rate of breaking Si-H bonds at the Si/SiO2 interface.

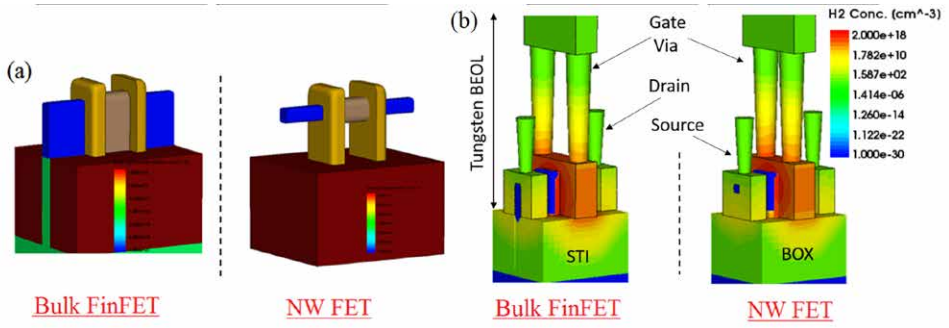


Figure 9. (Left) 3-D isometric view of the bulk FinFET and GAA NWFET structure used for simulation of advanced technology nodes. (right) Complete BEOL and H2 diffusion profile shown for the same structures at VOV = -0.85V, T = 1300C and tstress = 1000s. The Low-*k* material surrounding the gate metal is not shown for better viewing.

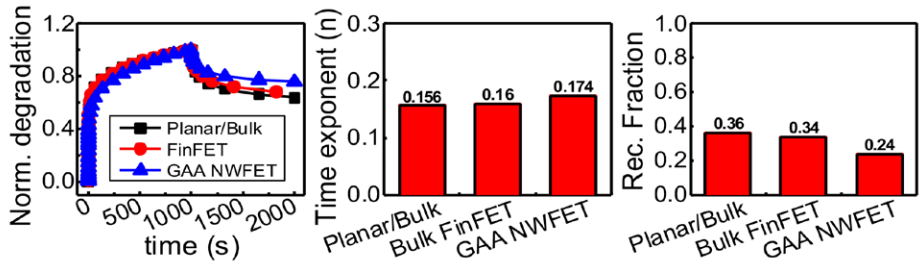
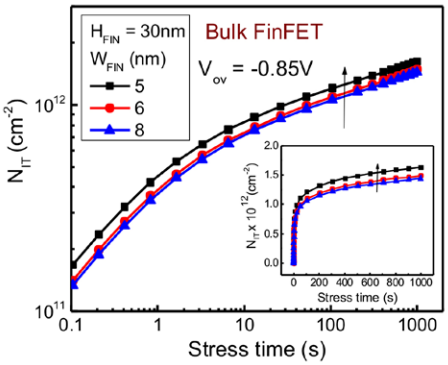


Figure 10. (Left) Time evolution of normalized degradation during stress and recovery for planar, FinFET and GAA NWFET. (right) Time exponent evaluated from 100s-1000s during stress for different architectures. (c) Recovery fraction for different architectures.

Similar longer time n is obtained due to identical backend for H_2 diffusion, and the fractional recovery is also similar across different device architectures. Note that GAA NWFET has slightly higher n and a slightly lower recovery fraction compared to planar and FinFET devices due to radial geometry [7].



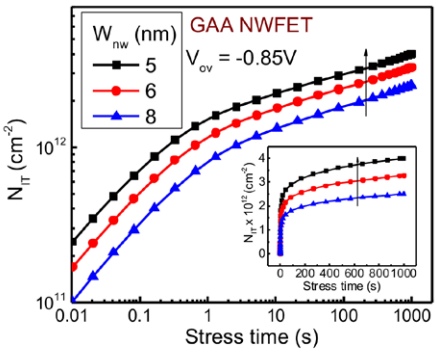


Figure 11. Simulated time evolution of ΔNIT for (top) bulk FinFET and (bottom) GAA NWFET with varying widths at VOV = -0.85V and Tstr = 1300C. Degradation increases with reducing width.

Scaling Impact

Scaling of device dimensions in advanced FinFET and GAA NWFET technology nodes requires the fins to be taller and thinner and nanowire widths to be smaller for better sub-threshold characteristics. In such ultra-scaled dimensions (below 10nm), both electrostatic and geometric quantum confinement effects become important and impact NBTI degradation.

Figures.11 (top) and (bottom) show the simulated time evolution of ΔVT (=ΔVIT) for different fin width (WFIN) and nanowire width (WNW) at constant overdrive (VOV) stress for bulk FinFETs and NWFETs respectively. It should be remarked that lower device widths result in higher trap density in both architectures. It is to be noted that the initial threshold voltage (VT0) increases due to increasing quantum confinement (QC) at lower widths requires increase in corresponding stress gate bias in such cases. This result can be attributed to higher surface hole density (p_s) and higher oxide electric field (EOX) at iso-VOV conditions, respectively shown in Figure 12 (left) and (right), which increase the forward reaction rate for Si-H bond dissociation as fin or NW widths are scaled. These have important implications on the device architecture changes and dimensional scaling on NBTI lifetime, refer to [6] for additional details.

Conclusion

A complete TCAD framework is developed to simulate interface trap generation and hole trapping in the gate insulator during NBTI stress in p-channel MOSFETs. The framework can predict measured NBTI stress and recovery quite accurately for different experimental conditions. The framework is employed to estimate NBTI degradation in bulk FinFETs and NWFETs having different fin and nanowire dimensions, to explore the impact of technology scaling on NBTI reliability.

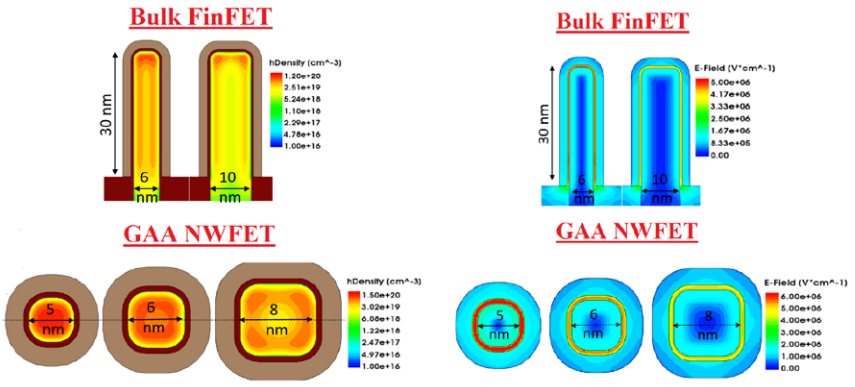


Figure 12. Simulated (left) inversion hole density and (right) electric field profile at a cross sectional plane perpendicular to the channel and located at the middle of the channel for bulk FinFETs and NWFETs, for varying W. VOV = -0.85V for all cases.

References

[1] S. Mahapatra, N. Goel, S. Desai, S. Gupta, B. Jose, S. Mukhopadhyay, K. Joshi, A. Jain, A. E. Islam, and M. A. Alam, "A comparative study of different physics-based nbti models," IEEE Trans. Electron Devices, vol. 60, no. 3, pp. 901–916, Mar. 2013.

[2] User's manual, Sentaurus TCAD.

[3] N. Parihar, N. Goel, A. Chaudhary, and S. Mahapatra, "A modelling framework for nbti degradation under dynamic voltage and frequency scaling," IEEE Trans. Electron Devices, vol. 63, no. 3, pp. 946–953, Mar. 2016.

[4] Tibor Grasser, "Stochastic charge trapping in oxides: From random telegraph noise to bias temperature instabilities," Microelectronics Reliability, Volume 52, Issue 1, January 2012, Pages 39-70.

[5] S. Mahapatra, V. Huard, A. Kerber, V. Reddy, S. Kalpat and A. Haggag, "Universality of NBTI - From devices to circuits and products", IEEE International Reliability Physics Symposium (IRPS), Kona, HI, USA, pp.3B.1.1-3B.1.8, 2014.

[6] S. Mishra; H. Y. Wong; R. Tiwari; A. Chaudhary; R. Rao; V. Moroz; S. Mahapatra, "TCAD-Based Predictive NBTI Framework for Sub-20-nm Node Device Design Considerations," in IEEE Transactions on Electron Devices, vol.PP, no.99, pp.1-8.

[7 H. Kufluoglu and M. A. Alam, "A geometrical unification of the theories of nbti and hci time-exponents and its implications for ultra-scaled planar and surround-gate mosfets," in Proc. IEEE Int. Electron Device Meeting, Dec. 2004, pp. 113–116.

SYNOPSYS®
Silicon to Software™

Synopsys, Inc. • 690 East Middlefield Road • Mountain View, CA 94043 • www.synopsys.com

©2016 Synopsys, Inc. All rights reserved. Synopsys is a trademark of Synopsys, Inc. in the United States and other countries. A list of Synopsys trademarks is available at http://www.synopsys.com/copyright.html. All other names mentioned herein are trademarks or registered trademarks of their respective owners.

11/23/16.CS7807_TCAD Newsletter.